Testing for linkage and Hardy-Weinberg disequilibrium

E KULINSKAYA[*], A LEWIN[†]

SUMMARY

This paper concerns several important points when testing for Hardy-Weinberg equilibrium (HWE) and linkage disequilibrium (LD) in genetics. First, we challenge the necessity of using exclusively two-sided tests for LD. Next, we show that the exact 2-sided tests based on the most popular measures of LD are not equivalent, and neither are the standard statistical tests even though the 1-sided tests are equivalent. We show how this results in different inference about LD for two data sets consisting of small groups of markers. Finally, we advocate the use of the conditional p-value for both LD and HWE testing. An important advantage of this p-value is that equivalent 1-sided tests are transformed into equivalent 2-sided tests.

[*] Statistical Advisory Service, Imperial College, South Kensington Campus, 8 Princes Gardens London SW7 1NA, UK. e-mail: e.kulinskaya@ic.ac.uk

[†] Dept. of Epidemiology and Public Health, Imperial College, St Mary's Campus Norfolk Place London W2 1P, UK. e-mail: a.m.lewin@imperial.ac.uk

keywords: nominal variables, Fisher's exact test, chi-square test, likelihood ratio test, Haldane test, measures of linkage disequilibrium

1 Introduction

Testing for linkage disequilibrium (LD) and Hardy Weinberg equilibrium (HWE) both involve the apparently simple problem of testing for independence in 2×2 contingency tables. However, there are many different possible test statistics commonly used. Hedrick (1987), Devlin & Risch (1995) and Mueller (2004) consider several different LD measures and their properties. Maiste & Weir (1995) compare tests for HWE using different test statistics.

In addition, all statistical tests for LD and HWE involve discrete test statistics and asymmetric null distributions (the hypergeometric for LD and Haldane for HWE). There is ongoing controversy about how 2-sided p-values should be constructed for the hypergeometric distribution (Agresti, 2002) and other non-symmetric distributions (Kulinskaya, 2008). Fisher advocated doubling the 1-sided p-value in his letter to Finney in 1946 (Yates, 1984, p.444), motivated by equal prior weights of departure in either direction. This choice has the drawback that the p-value can exceed 1. A popular two-sided p-value for non-symmetric discrete distributions implemented in computer packages is found by summing the probabilities of the points less probable than the observed (at both tails). Another possibility is to order the points by the squared or absolute values of correlation or other measures, when marker alleles are arbitrarily labelled (Mueller, 2004).

A new proposal for a two-sided p-value called the '*conditional p-value*' was introduced in Kulinskaya (2008). This new two-sided p-value has properties which make it a definite improvement on currently used two-sided p-values for both discrete and continuous non-symmetric distributions. It is closely related to the doubled p-value (but is automatically less than or equal to 1) and has an intuitive appeal.

In this paper we consider three kinds of p-values for exact 2-sided tests of LD: (1) the p-values which are the sum of the probabilities of the points less probable than observed (we refer to this as the standard 2-sided Fisher's test), (2) the p-values based on absolute values of different LD measures and test statistics and (3) the new conditional p-values.

The structure of the paper is as follows. In Section 2 we formulate the problems of linkage disequilibrium and linkage analysis and introduce some popular measures of LD. In Section 3 we point out that it can be appropriate to use one-sided tests; the use of two-sided tests is not required by invariance to relabelling of alleles. We also demonstrate that all measures of LD considered result in equivalent exact one-sided tests equal to the Fisher's exact test.

It is still more common to use 2-sided tests. In Section 4 we show that 2-sided p-values using the absolute values of different LD measures are *not* equivalent, and all are different to the Fisher's test and to the exact likelihood ratio test. Thus the choice of an appropriate 2-sided test for LD should depend on the measure of interest to a researcher. However, using the conditional p-values, all LD measures result in equivalent tests. This resolves the necessity of the careful choice of the 2-sided test for non-symmetric distributions. In Section 5 we show how the different LD tests result in different inference for two data sets consisting of small groups of markers.

In Section 6 we show that the same problems found for LD testing are also relevant for HWE testing, and propose the use of the conditional p-value with the Haldane test. Discussion is in Section 7.

A software package for R (R Development Core Team, 2004) is available from http://www.bgx.org.uk/alex/ or from the CRAN website http://cran.r-project.org/.

2 Measures of Linkage Disequilibrium

For bi-allelic markers at loci A and B, linkage disequilibrium data can be presented in the form of 2x2 contingency tables where haplotypes are classified in terms of their alleles at each of the 2 loci:

$$T_{AB} = \begin{array}{|c|cc|c|} & B_1 & B_2 & \text{total} \\ \hline A_1 & n_{11} & n_{12} & n_{1+} \\ A_2 & n_{21} & n_{22} & n_{2+} \\ \hline \text{total} & n_{+1} & n_{+2} & n \end{array}$$

Testing for linkage equilibrium is equivalent to testing for independence in the 2x2 table. Since there are often tables with low cell counts, the approximations used in the standard chi-squared test or in the likelihood ratio test are not valid, thus exact tests using the hypergeometric distribution are generally most appropriate.

Fisher's exact test uses n_{11} as the test statistic. Since the exact tests are conditional on the observed margins of the table (n_{1+}, n_{+1}, n), the n_{11} value defines all the other entries in the table. We shall refer to the respective haplotypic probabilities as p_{ij}, $i, j = 1, 2$, corresponding to the counts n_{ij}. These probabilities can be estimated by the observed haplotypic frequencies $\hat{p}_{ij} = n_{ij}/n$. A parameter of primary importance is the odds ratio $\rho = p_{11}p_{22}/p_{12}p_{21}$ estimated by $\hat{\rho} = n_{11}n_{22}/n_{12}n_{21}$. The case of no association $p_{ij} = p_{i+}p_{+j}$ is equivalent to $\rho = 1$.

Contingency tables for LD are often summarised by a measure of the degree of disequilibrium. The difference between the observed and expected frequencies is the *LD parameter* $D = p_{11} - p_{1+}p_{+1} = p_{22} - p_{2+}p_{+2} = -(p_{12} - p_{1+}p_{+2}) = -(p_{21} - p_{2+}p_{+1}) = p_{11}p_{22} - p_{12}p_{21}$. There exist a variety of disequilibrium measures, many of which are based on D standardized in different ways. Devlin & Risch (1995) discuss 5 popular measures listed here.

To reduce the dependence of D on allele frequencies Lewontin (1964) introduced $D' = D/D_{max}$, where D_{max} is the maximum value given the allele frequencies calculated as $D_{max} = \min\{p_{1+}p_{+2}, p_{2+}p_{+1}\}$ when $D > 0$, and $D_{max} = \min\{p_{1+}p_{+1}, p_{2+}p_{+2}\}$ when $D < 0$. The Pearson's correlation coefficient is $r = D/(p_{1+}p_{+1}p_{2+}p_{+2})^{1/2}$. Other popular measures include the difference in proportions $d = p_{11}/p_{+1} - p_{12}/p_{+2} = D/(p_{+1}p_{+2})$, an approximation for the population attributable risk under case-control sampling $\delta = D/p_{+1}p_{22}$, and Yule's $Q = (\rho - 1)/(\rho + 1) = D/(p_{11}p_{22} + p_{12}p_{21})$ (Devlin & Risch, 1995). The most frequently used measures of LD are D' and r.

Mueller (2004) further discusses the differing properties of D' and r^2 resulting in differing applications: D' is useful for assessing historical recombination in a given population, and the r^2 is useful in the context of association studies. Devlin & Risch (1995) proclaim δ to be superior for fine mapping because it is directly related to the recombination fraction.

Exact tests based on these different LD measures all use the hypergeometric distribution. Each possible 2x2 table has a particular probability under the null. The differences in 2-sided p-values come from the different orderings of the possible tables according to absolute value of LD measure.

3 One-sided statistical tests for Linkage Disequilibrium

The distribution of n_{11} is the hypergeometric (Fisher, 1935):

$$f(n_{11}; n_{1+}, n_{+1}; \rho) = \frac{\binom{n_{1+}}{n_{11}}\binom{n-n_{1+}}{n_{+1}-n_{11}}\rho^{n_{11}}}{\sum_u \binom{n_{1+}}{u}\binom{n-n_{1+}}{n_{+1}-u}\rho^u},$$

where ρ is the odds ratio. The null distribution (standard hypergeometric) has $\rho = 1$. For the one-sided test of $H_0 : \rho = 1$ vs $H_1 : \rho > 1$, Fisher proposed the p-value $p_+ = \sum_{u \geq n_{11}} f(u; n_{1+}, n_{+1}; 1)$, and for $H_0 : \rho = 1$ vs $H_1 : \rho < 1$ the p-value $p_- = \sum_{u \leq n_{11}} f(u; n_{1+}, n_{+1}; 1)$. This is known as the Fisher's Exact Test.

The one-sided p-values for tests using an LD measure L are similarly $p_+ = \sum_{L(u) \geq L(n_{11})} f(u; n_{1+}, n_{+1}; 1)$ and $p_- = \sum_{L(u) \leq L(n_{11})} f(u; n_{1+}, n_{+1}; 1)$. Here we consider 5 possible LD measures: $\hat{D}'$, $\hat{r}$, $\hat{\delta}$, $\hat{d}$, and $\hat{Q}$.

To be able to deduce the properties of the various measures of LD, let us relabel the probabilities as $p_{11} = x, p_{1+} = a, p_{+1} = b$ (Crook & Good, 1982). Then $D = x - ab$ and

$$\begin{array}{ll} D' & = (x-ab)/\min(a(1-b),(1-a)b) \text{ when } D > 0 \text{ and} \\ D' & = (x-ab)/\min(ab,(1-a)(1-b)) \text{ when } D < 0; \\ r & = (x-ab)/\sqrt{ab(1-a)(1-b)}; \\ d & = (x-ab)/(b(1-b)); \\ \delta & = b^{-1}(1-(1-a-b+ab)/(1-a-b+x)); \\ \rho & = 1+(x-ab)/((a-x)(b-x)); \\ Q & = 1-2/(1+\rho). \end{array} \qquad (1)$$

It is easy to see that all these functions are increasing functions of x, and all the resulting test statistics are increasing functions of n_{11} (note that $\hat{x} = n_{11}/n$). Therefore all the 1-sided tests are equivalent to Fisher's exact test. This was shown first by Davis (1986) for $\hat{r}$, $\hat{d}$ and $\hat{\rho}$. Thus Fisher's exact test is an appropriate 1-sided test to test that any of these measures are positive ($D > 0$) or negative ($D < 0$). Also Fisher's exact test is the Uniformly Most Powerful Unbiased (UMPU) test if the randomization is allowed (Tocher, 1950).

3.1 Invariance to relabelling

An exact test for association of two nominal variables should be invariant under relabelling of rows and columns. A two-sided version of Fisher's test is a traditional remedy when an invariance in respect to row/column relabelling is required. Is it appropriate?

Let us explore the effects of relabelling the rows of the table T_{AB}. The resulting table is

$$T_{\pi(A)B} = \begin{array}{|c|cc|c|} & B_1 & B_2 & \text{total} \\ \hline A_2 & n_{21} & n_{22} & n_{2+} \\ A_1 & n_{11} & n_{12} & n_{1+} \\ \hline \text{total} & n_{+1} & n_{+2} & n \end{array}.$$

The odds ratio is now $\tilde{\rho} = p_{21}p_{12}/p_{22}p_{11} = \rho^{-1}$ and $\tilde{D} = p_{21} - p_{2+}p_{+1} = -D$. Therefore all signed measures of LD change their sign, but are otherwise unchanged, except for δ, which is not invariant as it requires specification of case or control status.

The Fisher's test statistic after relabelling is $n_{21} = n_{+1} - n_{11}$, and it is easy to see that the probability $f(n_{21}; n_{2+}, n_{+2}; \tilde{\rho}) = f(n_{11}; n_{1+}, n_{+1}; \rho)$. Thus, the two tables have the same probability $P(T_{AB}) = P(T_{\pi(A)B})$. A 1-sided test for $\rho > 1$ is transformed into an *equivalent* test for $\tilde{\rho} < 1$, *i.e.* the p-value is invariant under the permutation of rows. This invariance of the p-value also applies to the relabelling of columns, and to changing the rows to columns and vice versa. Therefore Fisher's one-sided test is in fact invariant to relabelling, and thus is a valid test for association on a nominal scale.

The usual perception of a one-sided test is that it tests for a particular direction, say for $\rho > 1$. Given a particular labelling on A and B, a resulting sign of ρ merely indicates a prevalence of a particular combination of A and B values. This information does not change with relabelling, even though the sign of ρ and D does, thus the significance of LD (and p-value) does not change either. Therefore, when it is known which allele of a marker is associated with a disease, a one-sided test should be used. This is the case in confirmatory studies, for example, a candidate-gene study of a disease based on a different population. A two-sided test should be used when there is no knowledge of which allele in a marker is detrimental and which is protective to a disease.

4 Two-sided tests for Linkage Disequilibrium

In this section we consider six different exact 2-sided tests: the standard 2-sided Fisher's test, the exact test based on the likelihood ratio (LR) test statistic, tests using absolute values of the correlation coefficient, $|D'|$ and Yule's $|Q|$, and the conditional p-values introduced by (Kulinskaya, 2008). The exact tests based on the standard chi-squared statistic and on d are included in this comparison, as these are equivalent to the test based on the correlation coefficient. We do not look at 2-sided p-values based on $|\delta|$, as these are not invariant to row/column relabelling. For the purposes of testing independence between two loci this is not appropriate.

4.1 Table orderings for different LD measures

To calculate the exact two-sided tests for LD based on the absolute values of LD measures, we need to order all possible tables T_{AB} with given margins according to an LD measure of choice, calculate their probabilities using the hypergeometric distribution, and their p-values as the cumulative probabilities under the ordering. Thus in order to see the differences between the LD measures, we compare the orderings corresponding to each of the absolute values of the LD measures. The standard 2-sided Fisher's test uses the statistic $F_P = -P(n_{11})$, which orders the tables according to their probability.

Here we also consider the Likelihood Ratio test statistic $LR = 2\sum_{ij} n_{ij} \log(n_{ij}/m_{ij})$, where $m_{ij} = n_{i+}n_{+j}/n$ is the expected value of n_{ij} under the null hypothesis of no association, and $0 \times \log(0) = 0$ by definition (Agresti, 2002). The LD measure $|d|$ results in the same ordering of 2x2 tables as does $|r|$, and thus is omitted from the comparison. Note that the Pearson's chi-square test statistic $X^2 = \sum_{ij}(n_{ij} - m_{ij})^2/m_{ij} = nr^2$ is also equivalent to $|r|$, and so the exact version of this test is implicitly included in our comparison.

All five statistics F_P, LR, $|r|$, $|D'|$ and $|Q|$ are strictly decreasing functions of n_{11} for $n_{11} \leq m_{11}$ or equivalently for the LD parameter $D \leq 0$ ('the left tail') and strictly increasing functions of n_{11} for $n_{11} \geq m_{11}$ or $D \geq 0$ ('the right tail'), (Davis, 1986). The 2-sided test based on a statistic Y rejects for large values of $|Y|$, and the 2-sided p-value is calculated as $p(|Y|) = P(|y| \geq |Y|)$.

Example 1: Consider the table with margins $(n_{1+}, n_{2+}, n_{+1}, n_{+2}) = (9, 21, 5, 25)$ used as an example in Davis (1986). The possible n_{11} values are 0 through 5, the expected value is $m_{11} = 1.5$, so the left tail has two tables only, for $n_{11} = 0$ and 1, with the total probability of $w_L = 0.521$. Tables with $n_{11} = 2, \cdots, 5$ are on the right tail, the total probability is $w_R = 0.479$. The 6 tables, their exact probabilities based on hypergeometric distribution, and the respective values of 5 statistics of interest are given in Table 1. The values of F_P omitted from Table 1 to avoid duplication are easily obtained as $F_P = -P(T_{AB})$. Each table T_{AB} is uniquely defined by the value of n_{11}, and we shall refer to them by this number from now on.

n_{11}	n_{12}	n_{21}	n_{22}	$P(T_{AB})$	LR	$\|r\|$	$\|D'\|$	$\|Q\|$
0	9	5	16	0.143	1.990	0.293	1.000	1.000
1	8	4	17	0.378	0.152	0.098	0.333	0.306
2	7	3	18	0.336	0.137	0.098	0.143	0.263
3	6	2	19	0.124	1.184	0.293	0.429	0.652
4	5	1	20	0.019	3.314	0.488	0.714	0.882
5	4	0	21	0.001	7.334	0.683	1.000	1.000

Table 1: The 6 possible tables T_{AB} with margins $(n_{1+}, n_{2+}, n_{+1}, n_{+2}) = (9, 21, 5, 25)$ are given in columns 1-4, their probabilities $P(T_{AB})$ and the values of various LD measures are given in columns 5-9.

The tables are ordered by the increasing values of test statistics, as follows:

$$\begin{array}{ll} F_P: & 1\ 2\ 0\ 3\ 4\ 5 \\ LR: & 2\ 1\ 3\ 0\ 4\ 5 \\ |r|: & \{1\ 2\}\ \{0\ 3\}\ 4\ 5 \\ |D'|: & 2\ 1\ 3\ 4\ \{0\ 5\} \\ |Q|: & 2\ 1\ 3\ 4\ \{0\ 5\} \end{array}$$

Table 2 gives the p-values from the different orderings. Results for $|\delta|$ and $|Q|$ are omitted, as they coincide with those for LR and D' respectively. The last column provides the conditional p-values p_C discussed in the next Section. Only tables 4 and 5 have small enough probabilities to ever result in small p-values. The three standard tests (P_F, LR and the chi-square test based on $|r|$) have the largest test statistic values for tables 4 and 5. The p-values for all these three tests are $p(4) = 0.019$ and $p(5) = 0.001$, resulting in the conclusion of LD when one of these tables is observed. However a test based on $|D'|$ or $|Q|$ would result in $p(5) = p(0) = 0.144$ and $p(4) = 0.162$, thus for these two tests, neither table would draw a conclusion of LD.

$n11$	$n12$	$n21$	$n22$	p_{Fish}	p_{LR}	p_r	$p_{D'}$	p_C
0	9	5	16	0.286	0.162	0.286	0.144	0.274
1	8	4	17	1.000	0.664	1.000	0.664	1.000
2	7	3	18	0.622	1.000	1.000	1.000	1.000
3	6	2	19	0.143	0.286	0.286	0.286	0.299
4	5	1	20	0.019	0.019	0.019	0.162	0.041
5	4	0	21	0.001	0.001	0.001	0.144	0.002

Table 2: 2-sided p-values for linkage disequilibrium, example 1.

In this example the orderings based on D' and $|Q|$ coincide. This is not true in general, as will be seen in Section 4.3 for an example with larger sample size.

These results show that when using standard 2-sided tests, the choice of LD measure can have a large effect on inference. It makes sense to use a two-sided test based on the LD measure of interest. Only this can guarantee a consistency between the conclusions of the test and the degree of LD as given by the measure of choice. Our R package includes the calculation of all five 2-sided tests.

4.2 Conditional p-values for LD

The conditional 2-sided p-value $p_C(x)$ (Kulinskaya, 2008) is the one-sided p-value for the observed tail, conditioned on the observed tail. Effectively, the one-sided p-value is weighted by the probability of the tail. (Compare this to the doubled one-sided p-value, which weights always by 0.5.) For a continuous symmetric distribution the conditional p-value is the doubled p-value.

When calculating the conditional p-value, the first task is to decide on the point A separating the two tails of the distribution in question. The mean or expected value E is often the most suitable choice, but it may be instead the median m, the mode M, or some other location parameter.

Definition 4.2.1 *Conditional p-value for a discrete distribution is*

$$p_C(x|A) = \begin{cases} P(X \leq x)/w_L, & x < A; \\ 1 & x = A; \\ P(X \geq x)/w_R, & x > A; \end{cases} \tag{2}$$

where the weights are $w_L = P(x \leq A)$ *and* $w_R = P(x \geq A)$.

When $A = m$ is an attainable value of the discrete distribution, the values of $p_C(x|m)$ are $(1 + P(m))$ times smaller than doubled 1-sided p-values. When $A = E$ as is often more appropriate, the weights of the tails differ unless the distribution is symmetric. The conditional p-value has a mode of 1 at A when this value is attainable, and two modes of 1 at the attainable values above and below A when A is not an attainable value.

For LD testing we use the choice $A = E = m_{11}$, which means that the tail is defined by the sign of the LD parameter D, similar to the previous Section.

The critical region for any two-sided test at level α is defined by probabilities $\alpha_1 = w_L\alpha$ and $\alpha_2 = w_R\alpha$, with the weights of the two tails $w_L + w_R = 1$. The two-sided test corresponding to the conditional p-value corresponds to the choice of weights w_L and w_R as in definition 4.2.1.

Lemma 1 from Kulinskaya (2008) ensures that equivalent 1-sided tests are transformed into equivalent 2-sided tests when the conditional p-value $p_C(x|E)$

is used. This is true because the conditional p-value ignores any equivalence between the points at different tails. Therefore all five statistics F_P, LR, $|r|$, $|D'|$ and $|Q|$ discussed in Section 4.1 for LD testing result in the same 2-sided tests when the conditional p-value is used (as we saw in Section 3, these all give equivalent 1-sided tests).

The conditional p-values are included in Table 2 for the Example in the previous Section. The table with $n_{11} = 0$ is on the left tail and $p_C(0|E) = 0.143/0.521 = 0.274$. The tables with $n_{11} = 4$ and 5 are on the right tail, and the p-values are $p_C(4|E) = 0.019/0.479 = 0.041$ and $p_C(5|E) = 0.001/0.479 = 0.002$. The conclusions coincide with those from the three standard tests but the p-values are noticeably larger.

4.3 Large sample behaviour of exact tests for LD

The differences between p-values obtained using different LD measures remain for large sample sizes. Figure 1 shows the two-sided p-values from the three standard tests (Fisher's, Likelihood ratio and correlation-based) and the conditional p-values for two different null distributions with large sample sizes ($n = 500$ and $n = 1000$). It is clear that there are still considerable differences between the p-values from different tests. In some cases this would lead to different conclusions being drawn from the different tests.

The null distributions used for the illustration here have the same ratios of margins to sample size $n_{1+} : n_{+1} : n$, but different tests give larger p-values in the two cases. We have not been able to discern a pattern in the behaviour of the p-values from these four tests as sample size increases.

Figure 2 shows the p-values based on D' and Yule's Q, with Fisher's p-values for comparison, for two null distributions with sample size 1000. Here there are very large differences between the p-values. In the very skewed case, $(n_{1+}, n_{+1}, n) = (20, 50, 1000)$, the D' and Q p-values can never be small enough for any observation to reject the null hypothesis, though there are tables with very small probabilities under the null, which would lead to rejection of the null hypothesis if Fisher's test were used. In the less skewed case, $(n_{1+}, n_{+1}, n) = (130, 150, 1000)$, the p-values from Yule's Q are closer to the Fisher's p-values, but the D' p-values still show very large differences.

The D' and Yule's Q p-values become closer to the Fisher's p-values as sample size increases, but for any given sample size there will always be null distributions too skewed to be rejected regardless of observations using the D' and Q tests.

The reason for the large $|D'|$ and $|Q|$ p-values in the skewed case is because both these statistics are scaled in such a way that the tables at the two extremes (those with the smallest and largest values of n_{11}) have statistics with absolute value close to 1. This scaling means that the tables at the two extremes are effectively given similar weights in the hypothesis tests. Hence the tables with small n_{11}, which in the skewed case have high probability under the null, contribute to the p-values for tables with large n_{11}. This results in large p-values for tables with large n_{11}, despite the fact that these tables actually have very small probabilities under the null hypothesis. These results reflect the fact that $|D'|$ is less powerful to detect LD when a rare

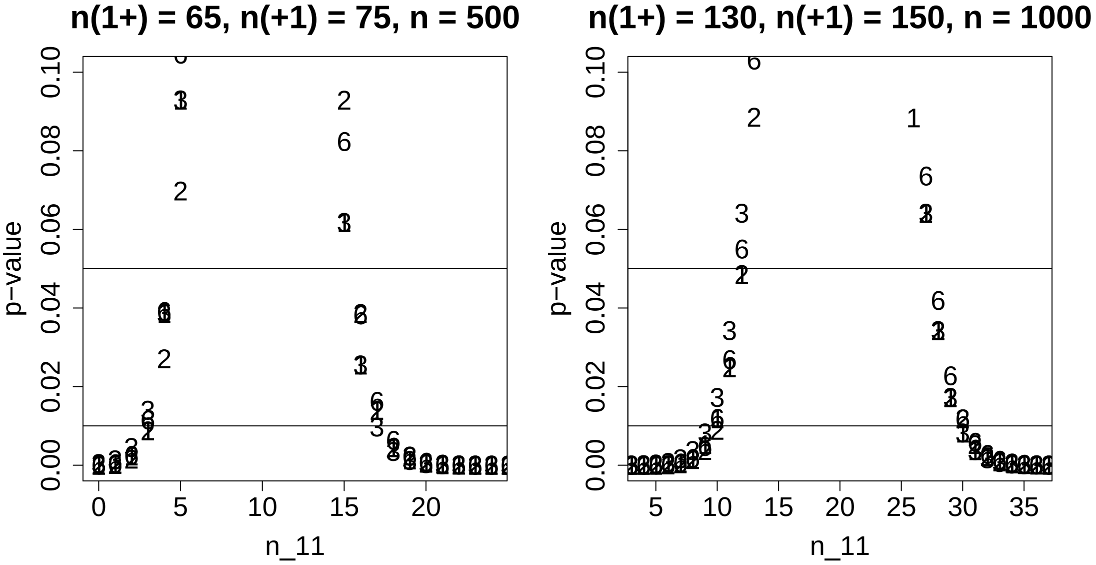

Figure 1: Two-sided p-values for the three standard LD measures and the conditional p-values, for all possible values of n_{11} under two different null distributions. Left: $(n_{1+}, n_{+1}, n) = (65, 75, 500)$. Right: $(n_{1+}, n_{+1}, n) = (130, 150, 1000)$. Symbols used are: 1 Fisher's p-values, 2 Likelihood ratio test p-values, 3 correlation-based p-values, 6 conditional p-values. The x-axes are limited to show the non-zero p-values. The y-axes are focused on small p-values to enable the differences to be seen. Lines at 0.05 and 0.01 indicate the thresholds traditionally used to assess significance.

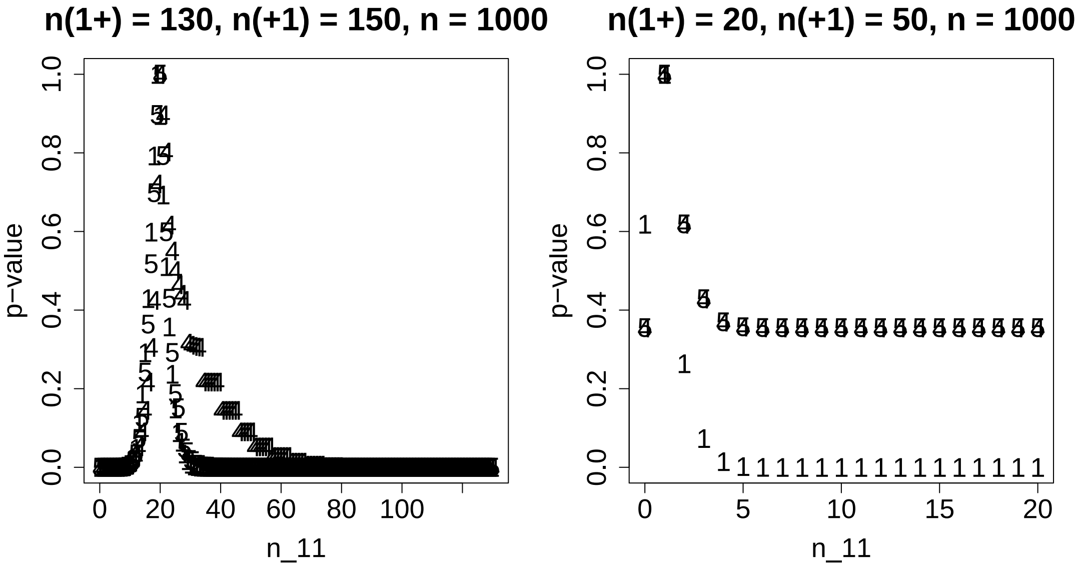

Figure 2: Two-sided p-values for the Fisher's test and for the tests based on D' and Yule's Q, for all possible values of n_{11} under two different null distributions. Left: $(n_{1+}, n_{+1}, n) = (130, 150, 1000)$. Right: $(n_{1+}, n_{+1}, n) = (20, 50, 1000)$. Symbols used are: 1 Fisher's p-values, 4 p-values based on D', 5 p-values based on Yule's Q.

allele is associated with another rare allele or a very common allele.

5 Application to genetic marker sets

We apply the tests for LD to two data sets. The first is a set of 28 2x2 tables resulting from pairwise comparisons of 8 RFLPs (restriction fragment length polymorphisms) at the insulin receptor (INSR) locus, obtained from 228 independent haplotypes (Elbein, 1992). Seven RFLPs were examined in all 228 haplotypes and an additional RFLP was included for 172 of the haplotypes. Thus 7 tables have sample size $n = 172$ and the remaining 21 have $n = 228$. The tables do not in general have the same margins, so the null distributions are different.

The second data set is another set of 28 comparisons of 8 RFLPs, this time at the phenylalanine hydroxylase locus, from 33 families with at least one phenylketonuric (PKU) child (Chakraborty et al., 1987). Independent haplotypes were obtained from 66 parents, 33 with the PKU mutation and 33 without. Thus the sample size n is 66 for all tables, again with different null distributions. PKU and non-PKU haplotypes were analysed separately.

Figure 3 shows the results of the different tests of linkage disequilibrium for the Elbein data set, controlling the false discovery rate at 5% using the Benjamini & Hochberg (1995) procedure. The lower panel shows the different p-values for the 28 tables, with the tables arranged along the x-axis in order of increasing Fisher's p-value. The upper panel indicates with crosses for

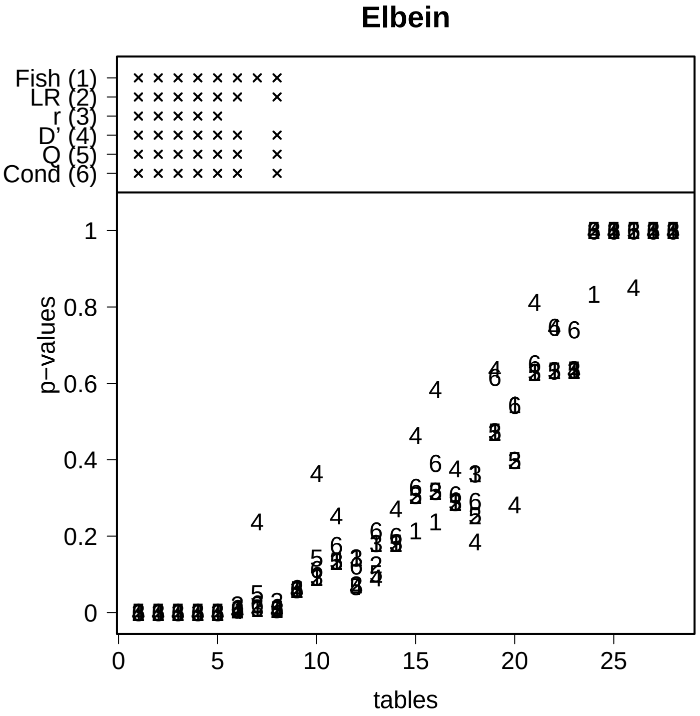

Figure 3: Six different LD tests for the Elbein data (28 2x2 tables). Upper panel indicates for which tables the null hypothesis of no LD is rejected, controlling the false discovery rate at 5% (crosses indicate rejection of the null). Lower panel shows the two-sided p-values from all tests. Symbols used are: 1 Fisher's p-values, 2 Likelihood ratio test p-values, 3 correlation-based p-values, 4 p-values based on D', 5 p-values based on Yule's Q, 6 conditional p-values. Tables are ordered by increasing Fisher's p-value.

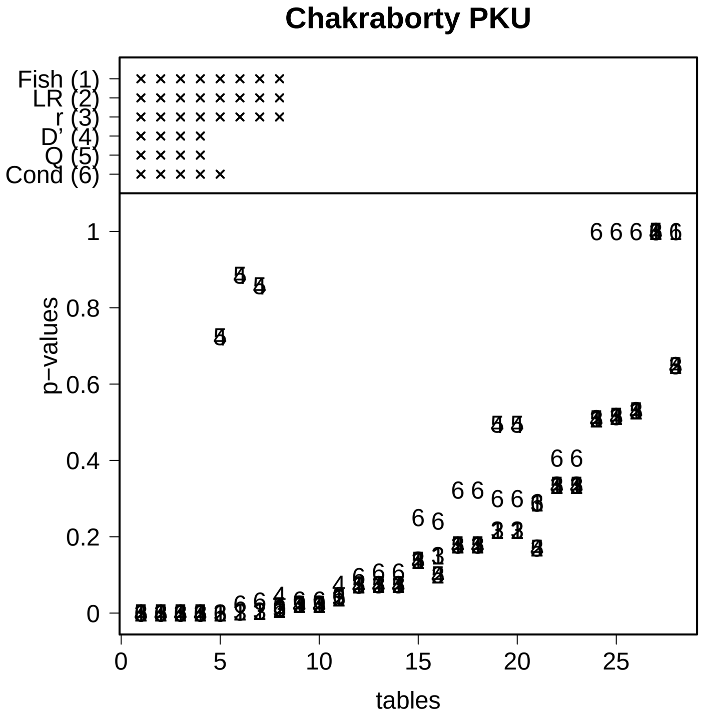

Figure 4: Six different LD tests for the Chakraborty data (PKU only). Plot format as Figure 3.

which tables the null hypothesis is rejected. In this case the three standard tests (Fisher's, Likelihood ratio and correlation-based) reject the null for different sets of tables. The conditional p-value provides the same results as the Likelihood ratio test, D' and Q, and the correlation-based test is the most conservative for this data.

Figures 4 and 5 show the equivalent plots for the Chakraborty data, with PKU and non-PKU analysed separately (as in the original work). In both cases the three standard tests give equivalent inference. The tests based on D' and Yule's Q give the same results as each other, and the conditional test

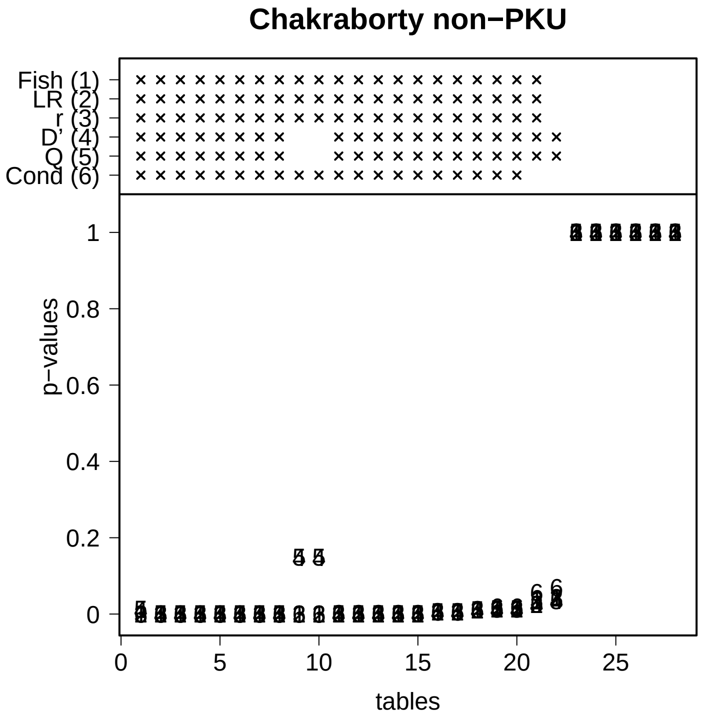

Figure 5: Seven different LD tests for the Chakraborty data (non-PKU only). Plot format as Figure 3.

is different again. In both cases it is more conservative than the standard three tests for this data.

All three data sets contain tables which have very large D' and Q p-values but small p-values from all other tests. This suggests that these two measures of LD should be used with great caution. There is no discernable pattern in the behavior of the tests based on r, LR, Fisher, or conditional p-values. In the Elbein data, P_C-based test coincides with the majority vote of the three standard tests; in the Chakraborty data it is the most conservative. In each data set, and in general, it does not seem obvious which out of r, Fisher and LR to choose; unified inference for the different LD measures based on the conditional p-value seems a reasonable choice.

6 Testing for Hardy-Weinberg equilibrium

In this section we consider the problem of testing for Hardy-Weinberg equilibrium (HWE). For a locus with two alleles A and a in a sample of size n genotypes, denote by n_{AA}, n_{Aa}, n_{aa} the observed genotypic counts. The gene frequencies are denoted by $n_A = 2n_{AA} + n_{Aa}$ and $n_a = 2n - n_A$, with $n_a < n_A$ (so $n_a/2n$ is the minor allele frequency). The probability of the observed set of heterozygotes $n_{Aa} = x$ is (Levene, 1949)

$$P(x|n_A) = \frac{n!n_A!(2n - n_A)!2^x}{[(n_A - x)/2]!x![n - (n_A + x)/2]!(2n)!}$$

A conditional exact test based on the fixed gene frequencies is named after Haldane (1954) though earlier publications include Levene (1949) and

Stevens (1938). We call this distribution the Haldane distribution. The sample set is the full set of possible number of heterozygotes x which is either even or odd numbers from 0 to n_a depending on whether n_A (and $n_a = 2n - n_A$) is even or odd. The distribution $P(x|n_A)$ is a unimodal non-symmetric distribution with the mode $(n_A n_a - 2)/(2n+3) \leq$ mode $\leq 2 + (n_A n_a - 2)/(2n+3)$ (Vithayasai, 1973). The expected values of heterozygotes under HWE is $\mathrm{E}(n_{Aa}) = n_A n_a/(2n-1)$ (Levene, 1949). A 1-sided test would reject for either small or large values of n_{Aa}, depending on whether inbreeding $(n_{Aa} < \mathrm{E}(n_{Aa}))$ or outbreeding $(n_{Aa} > \mathrm{E}(n_{Aa}))$ is the alternative of interest.

The standard exact 2-sided test for HWE is based on the ordering induced by $P(x|n_A)$. We denote the corresponding p-value by p_H. The distribution of $P(x|n_A)$ is asymptotically Normal, and the search for a suitable approximation to the exact test generated numerous contenders. A list of 10 asymptotic tests all based on the chi-square(1) distribution is given by Emigh (1980). The 2-sided tests result in differing orderings on the sample set, and provide quite different p-values, especially for the intermediate values of $n_a/(n_A + n_a) < 0.5$ (Emigh, 1980). Wigginton et al. (2005) demonstrated that the chi-square approximation results in inflated type 1 error rates in comparison to the exact 2-sided test even for large $n = 1000$ when $n_a = 100$. An efficient calculation of the exact test is given by Wigginton et al. (2005), cancelling the rationale of using the asymptotic tests. But the definition of the exact 2-sided test or the corresponding p-value is a problem very similar to that discussed for Fisher's exact test for LD in the previous section. We advocate the use of the conditional p-value $p_C(x|E)$ with the Haldane distribution. We denote this 2-sided p-value by $p_{HC}(x)$.

When $n_a < n_A$ the left tail of the Haldane distribution is considerably longer and somewhat heavier than the right tail, i.e. $w_L > w_R$. See Figure 6 for some examples. The smallest probability on the right tail (for $x = n_a$) can be rather large, and is always larger than the probabilities for a whole interval of small x values on the left tail. As a result the 2-sided p-value $p_H(x)$ coincides with the p-value for the 1-sided test of inbreeding for this range of small x values. This may lead to too many rejections. The 2-sided p_{HC}-based test is more conservative in this case than the standard p_H-based Haldane test and may show even more differences in the p-values with the chi-square test than those found by Wigginton et al. (2005) for the p_H test. This makes the exact calculation paramount.

The conditional p-value is easily calculated with a minor modification of the Wigginton et al. (2005) algorithm. This is included in our R package.

As an example, consider a case with number of genotypes $n = 100$ and $n_a = 34$ (minor allele frequency 0.17) given in Table 2 of Emigh (1980). The number of heterozygotes can be an even number from 0 to 34 (18 possible values). The null distribution is shown in Figure 6. The mean is 28.4, and the mode is an even number between 27.8 and 29.8, therefore equal to 28. The left tail consists of $\{2y, y \leq 14\}$, the weight is $w_L = 0.569$, and the right tail consists of only 3 values: 30, 32, and 34; the sum of the three probabilities is $w_R = 0.431$.

Table 3 shows the 1-sided inbreeding p-values (p_{in}), the 2-sided Haldane and conditional p-values. The probabilities in the right hand tail of the null distribution are all larger than the probability of $x = 22$, thus the 2-sided

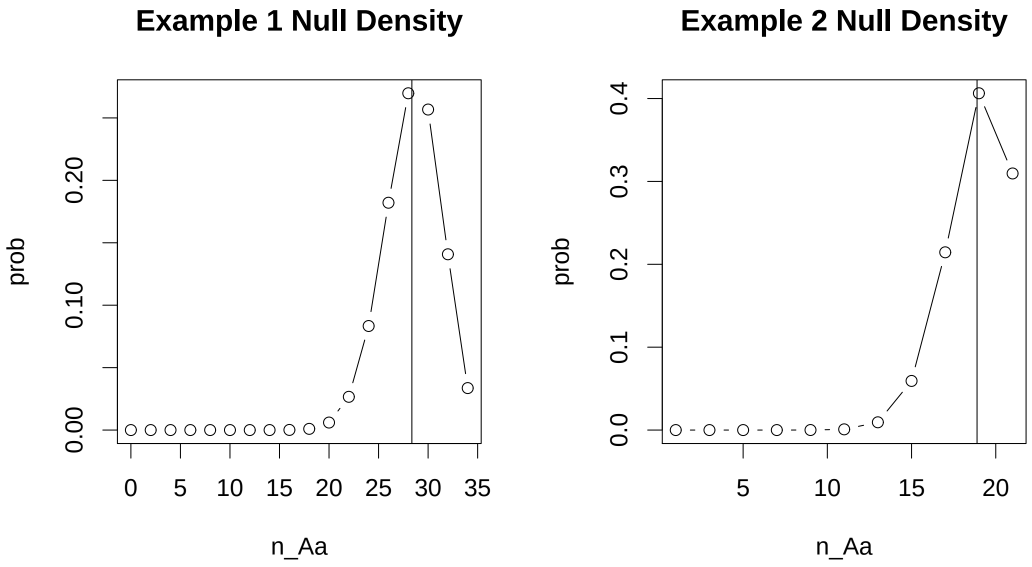

Figure 6: Null densities of n_{Aa} for the two examples. Vertical lines indicate the mean under the null.

n_{Aa}	$p_{in}(n_{Aa})$	$p_H(n_{Aa})$	$p_{HC}(n_{Aa})$
14	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
16	0.0001	0.0001	0.0002
18	0.0011	0.0011	0.0019
20	0.0071	0.0071	0.0125
22	0.0337	0.0337	0.0593
24	0.1171	0.1507	0.2058
26	0.2991	0.4735	0.5258
28	0.5689	1.0000	1.0000
30	0.8256	0.7303	1.0000
32	0.9664	0.2915	0.4045
34	1.000	0.0674	0.0780

n_{Aa}	$p_{in}(n_{Aa})$	$p_H(n_{Aa})$	$p_{HC}(n_{Aa})$
1	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
3	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
5	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
7	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
9	$< 10^{-4}$	$< 10^{-4}$	0.0002
11	0.0009	0.0009	0.0032
13	0.0103	0.0103	0.0362
15	0.0696	0.0696	0.2450
17	0.2840	0.2840	1.0000
19	0.6904	1.0000	1.0000
21	1.000	0.5936	0.4324

Table 3: **Left**: 1-sided inbreeding p-values, 2-sided Haldane and conditional p-values for the first example, for n_{Aa} between 14 and 34. Lower values of n_{Aa} have p-values of less than 10^{-4} in both cases. **Right**: Similar for the second example.

p_H values up to $x = 22$ coincide with the 1-sided p-values. In particular, $p_H(20) = P(x \leq 20) = 0.007$ and $p_H(22) = P(x \leq 22) = 0.034$. Using the conditional 2-sided test the p-values are almost doubled.

Next comes the most extreme point in the right hand tail, with probability $P(34|34) = 0.0336$ and 2-sided p_H p-value of $p_H(34) = 0.067$. This is the standard situation commented upon in Emigh (1980): the exact 2-sided test at $\alpha = 0.05$ level is equivalent to the 1-sided test for inbreeding, and cannot detect outbreeding. The latter cannot be helped due to the large probabilities on the right tail, but the lack of any penalty for using a 2-sided test instead of a 1-sided test for inbreeding seems wrong. If the conditional 2-sided test is used instead, the p-value is $p_{HC}(34) = 0.0336/0.431 = 0.078$.

An even more extreme example for $n = 100, n_a = 21$ (minor allele frequency 0.11) is considered by Wigginton et al. (2005). In their example the expected value $\mathrm{E}(n_{Aa}) = 18.89$, and there are only two points on the right tail with probabilities $P(19|21) = 0.406$ and $P(21|21) = 0.310$ (see Figure 1). The 2-sided p_H test is equivalent to the test for inbreeding for all points on the left tail $x \leq 17$ because the largest probability on the left tail is $P(17|21) = 0.214$. Since the weight of the left tail is $w_L = 0.284$, the p_{HC} values are 3.52 times larger.

As the sample size increases, the differences between the conditional p-values and the two-sided Haldane p-values grow smaller. However for genes with small minor allele frequency, there can still be differences for substantial sample sizes. For example, with 500 genotypes and a minor allele frequency of 0.05 ($n = 500, n_a = 50$) the conditional p-values for the left hand tail are approximately twice the Haldane p-values.

7 Discussion

The routine genetics problems such as testing for Hardy-Weinberg and linkage disequilibrium give rise to non-trivial statistical issues. This is due to the fact that the underlying distributions are discrete and non-symmetric.

We believe that the two-sided tests for LD and HWE are over- and misused. We showed that the usage of the two-sided tests is not necessitated by the invariance to relabelling. The 1-sided tests should be used when the direction of the association is known from prior research or is of particular interest. All 1-sided tests for LD are equivalent.

An important example is the fine-mapping of a disease-susceptibility locus. This can be achieved by testing for LD between disease and marker loci, or (for both recessive and additive disease model) by testing for HWE amongst cases only at a marker locus (Feder et al., 1996; Nielsen et al., 1999; Song & Elston, 2006). For recessive disease model, excess homozygosity conventionally indicating inbreeding indicates the proximity to the disease locus (Feder et al., 1996). In a general disease model, the direction of deviation from HWE is completely defined by the model (Nielsen et al., 1999; Wittke-Thompson et al., 2005; Zheng & Ng, 2008). Therefore for Mendelian diseases the direction of interest is usually known, and exact 1-sided tests are considerably more powerful than any two-sided tests, including the traditional chi-square test. The use of two-sided tests makes sense only for complex diseases where this direction may be unknown.

We showed the non-equivalence of the most popular 2-sided tests for LD, such as the Fisher's exact test, the exact chi-square and likelihood ratio tests, and the discrepancies in their results with those from the most popular measures of LD. An important conclusion of this paper is that a choice of a 2-sided test for LD should be based on a measure of interest to a researcher. To influence the practice, we provide the R package which calculates 6 exact tests based on the most popular measures of LD.

Two-sided statistical tests and p-values are well defined only when the test statistic in question has a symmetric distribution. Then the doubled 1-sided p-value makes perfect sense. But for non-symmetric distributions such as hypergeometric and Haldane distributions used in testing for LD or HWE, respectively, there is no consensus on how the 2-sided p-value should be defined.

We advocate the use of the conditional p-value introduced by Kulinskaya (2008) for both LD and HWE testing. This is the p-value given the tail. It weighs the tails inversely proportionate to their probabilities. In other words, it evaluates how unusual the observed value is given the direction of departure from the null hypothesis. It does not add up the probabilities of values at opposite direction. Given the importance of the direction for both LD and HWE, this has an intuitive appeal for a geneticist. An important advantage of this p-value bf from a statistical point of view is that equivalent 1-sided tests are transformed into equivalent 2-sided tests. When testing for LD, this means that all tests for LD provide the same results. For quality control, where markers are selected by comparing p-values to a threshold value, this unification means that the sets of markers selected are consistent, whichever LD measure is used. Our R package includes these conditional tests for both LD and HWE.

The tests considered in this paper are conditional exact tests. The distributions are conditional on the total gene frequencies. A different class of exact tests are unconditional tests. For LD, the conditional tests are based on the hypergeometric distribution; the most popular representative of this class is Fisher's exact test. Unconditional tests go back to Barnard (1947) and are, in general, less accepted. Unconditional tests may be more powerful and therefore require smaller sample sizes than the conditional tests (Suissa & Shuster, 1985), but their power largely depends on the chosen test statistic and a poor choice can result in a less powerfull analysis in comparison to conditional tests (Mehrotra et al., 2003). For LD exact unconditional tests for the difference in two binomial proportions implemented in StatXact (www.cytel.com) can be used. For the HWE an unconditional exact test based on the chi-square statistic is given in Haber (1994). Another test based on the Bayes factor was suggested by Montoya-Delgado et al. (2001). Both tests are two-sided by design. There also exists a considerable literature on the Bayesian methods in LD and HWE (Shoemaker et al., 1998; Sebastiani & Abad-Grau, 2007) among others. The effect of the choice of LD parameter is explicit in a Bayesian analysis, as this must be specified as part of a model. Additionally the choice of prior may affect the inference. The specific issue discussed in this paper regarding definition of the p-value arises from the different possible orderings of 2x2 tables that might be observed under the null hypothesis. This issue does not arise in Bayesian analysis as inference is conditional upon the observed 2x2 table.

Another important statistical issue, only mentioned in passing in Section 5, is the multiplicity of tests when testing for LD or HWE. Family-wise error rate procedures, such as Bonferroni, are much too stringent. False discovery rate (FDR) based procedures (Benjamini & Hochberg, 1995) are more suitable. An important advantage of the conditional p-value in this context is that it has discrete uniform distribution at each tail under the null hypothesis of equilibrium. This enables its use in fuzzy FDR procedures introduced for discrete distributions by Kulinskaya & Lewin (2008), resulting in the comprehensive statistical approach to LD and HWE testing. The finer details of this approach to multiple testing in genetics are to be described elsewhere.

Acknowledgements

We thank David Balding for helpful comments.

References

Agresti, A. 2002, Categorical Data Analysis, 2nd edn. (New York: John Wiley and Sons Ltd).

Barnard, G. (1947). Significance tests For 2X2 tables. *Biometrika* **34**, 123–138.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289–300.

Chakraborty, R., Lidsky, A. S., Daiger, S. P., Güttler, F., Sullivan, S., Diliella, A. G., and Woo, S. L. C. (1987). Polymorphic DNA haplotypes at the human phenylalanine hydroxylase locus and their relationship with phenylketonuria. *Human Genetics* **76**, 40–46.

Crook, J. F. & Good, I. J. (1982). The Powers and Strengths of Tests for Multinomials and Contingency Tables. *J. Am. Statist. Assoc.* **77**, 793–802.

Davis, L. (1986). Exact Tests for 2 ×2 Contingency Tables. *The American Statistician* **40**, 139–141.

Devlin, B. & Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322.

Elbein, S. C. (1992). Linkage Disequilibrium among RFLPs at the Insulin-Receptor Locus despite Intervening Alu Repeat Sequences. *Am. J. Hum. Gen.* **51**, 1103–1110.

Emigh, T. (1980). A Comparison of Tests for Hardy-Weinberg Equilibrium. *Biometrics* **36**, 627–642.

Feder, J., Gnirke, A., Thomas, W., Tsuchihasi, Z., Ruddy, D., Basava, A., and Dormishian, F. e. a. (1996). A novel MHC class Ilike gene is mutated in patients with hereditary haemochromatosis. *Nature Genetics* **33**, 399–408.

Fisher, R. A. (1935). The logic of inductive inference. *J. R. Statist. Soc. A* **98**, 39–54.

Haber, M. (1994). An Exact Unconditional test for the Hardy-Weinberg equilibrium. *Biometrical Journal* **36**, 741–749.

Haldane, J. (1954). An exact test for randomness of mating. *Journal of Genetics* **52**, 631–635.

Hedrick, P. W. (1987). Gametic Disequilibrium Measures: Proceed With Caution. *Genetics* **117**, 331–341.

Kulinskaya, E. (2008). On two-sided p-values for non-symmetric distributions. arXiv:0810.2124v1 [math.ST].

Kulinskaya, E. & Lewin, A. (2008). On fuzzy family wise error rate and false discovery rate procedures for discrete distributions. *Biometrika* in press.

Levene, H. (1949). On a matching problem arising in genetics. *Ann. of Math. Statist.* **20**, 91–94.

Lewontin, R. (1964). The integration of selection and linkage. I.General considerations; heteroic models. *Genetics* **49**, 49–67.

Maiste, P. J. & Weir, B. S. (1995). A comparison of tests for independence in the FBI RFLP data bases. *Genetica* **96**, 125–138.

Mehrotra, D., Chan, I., and Berger, R. (2003). A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics* **59**, 441–450.

Montoya-Delgado, L., Irony, T., de B. Pereira, C., and Whittle, M. (2001). Unconditional Exact Test for the Hardy-Weinberg Equilibrium Law: Sample-Space Ordering Using the Bayes Factor. *Genetics* **158**, 875–883.

Mueller, J. (2004). Linkage disequilibrium for different scales and applications. *Briefings in Bioinformatics* **5**, 355–364.

Nielsen, D., Ehm, M., and Weir, B. (1999). Detecting Marker-Disease Association by Testing for Hardy-Weinberg Disequilibrium at a Marker Locus. *Am. J. Hum. Genet.* **63**, 1531–1540.

R Development Core Team. 2004, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, iSBN 3-900051-00-3.

Sebastiani, P. & Abad-Grau, M. (2007). Bayesian estimates of linkage disequilibrium. *BMC Genetics* **8**, 36.

Shoemaker, J., Painter, I., and Weir, B. S. (1998). A Bayesian Characterization of Hardy-Weinberg Disequilibrium. *Genetics* **149**, 2079–2088.

Song, K. & Elston, R. (2006). A powerful method of combining measures of association and HardyWeinberg disequilibrium for fine-mapping in case-control studies. *Statistics in Medicine* **25**, 105–126.

Stevens, W. (1938). Estimation of blood group gene frequencies. *Annals of Eugenics* **8**, 377–383.

Suissa, S. & Shuster, J. (1985). Exact unconditional sample sizes for the 2 x 2 binomial trial. *Journal of the Royal Statistical Society, Series A* **148**, 317–327.

Tocher, K. (1950). Extension of the Neyman-Pearson Theory of Tests to Discontinuous Variates. *Biometrika* **37**, 130–144.

Vithayasai, C. (1973). Exact critical values of the Hardy-Weinberg test statistic for two alleles. *Communications in Statistics* **1**, 229–242.

Wigginton, J., Cutler, D., and Abecasis, G. (2005). A Note on Exact Tests of Hardy-Weinberg Equilibrium. *Am. J. Hum. Genet.* **76**, 887–893.

Wittke-Thompson, J. K., Pluzhnikov, A., and Cox, N. (2005). Rational inferences about departure from Hardy-Weinberg equilibrium. *American Journal of Human Genetics* **76**, 967–986.

Yates, F. (1984). Tests of significance for 2 × 2 contingency tables (with Discussion). *J. R. Statist. Soc. A* **147**, 426–463.

Zheng, G. & Ng, H. K. T. (2008). Genetic model selection in twophase analysis for case-control association studies. *Biostatistics* **9**, 391–399.